A framework for evaluating automatic indexing or classification in the context of retrieval

Koraljka Golub*

Department of Library and Information Science

School of Cultural Sciences

Faculty of Arts and Humanities

Linnaeus University

351 95 Växjö, Sweden

Tel: +46 (0) 470 708940

Fax: +46 (0) 470 751888.

E-Mail: koraljka.golub@lnu.se

*corresponding author

Dagobert Soergel

Department of Library and Information Studies

Room 534 Baldy

University of Buffalo

Buffalo, NY 14260-1020

United States of America

Tel: +1 716-645-1472

Fax: +1 3716 645-3775

E-Mail: dsoergel@buffalo.edu

George Buchanan

Department of Computer Science

City University

Northampton Square

London EC1V 0HB

United Kingdom

Tel: +44 (0)20 7040 8469

E-Mail: George.Buchanan.1@city.ac.uk

Douglas Tudhope

Hypermedia Research Group

University of South Wales

Pontypridd CF37 1DL

Wales, United Kingdom

Tel: +44 (0) 1443 483609

E-Mail: douglas.tudhope@southwales.ac.uk

Debra Hiom

IT Services R&D

University of Bristol

8-10 Berkeley Square

Bristol BS8 1HH

United Kingdom

Tel: +44 (0) 117 331 4381

E-Mail: d.hiom@bristol.ac.uk

Marianne Lykke

E-Learning Lab

University of Aalborg

Rensburggade 14

9000 Aalborg

Denmark

Tel: (+45) 9940 8157 /2125 1854

E-Mail: mlykke@hum.aau.dk

Table of contents

Abstract

Tools for automatic subject assignment help deal with scale and sustainability in creating and enriching metadata, establishing more connections across and between resources and enhancing consistency. While some software vendors and experimental researchers claim the tools can replace manual subject indexing, hard scientific evidence of their performance in operating information environments is scarce. A major reason for this is that research is usually conducted in laboratory conditions, excluding the complexities of real-life systems and situations. The paper reviews and discusses issues with existing evaluation approaches such as problems of aboutness and relevance assessments, implying the need to use more than a single "gold standard" method when evaluating indexing and retrieval and proposes a comprehensive evaluation framework. The framework is informed by a systematic review of the literature on indexing, classification and approaches: evaluating indexing quality directly through assessment by an evaluator or through comparison with a gold standard; evaluating the quality of computer-assisted indexing directly in the context of an indexing workflow, and evaluating indexing quality indirectly through analyzing retrieval performance.

Introduction

Subject terms play a crucial role in resource discovery but require substantial effort to produce. Automatic indexing addresses problems of scale and sustainability and can be used to enrich existing bibliographic records, establish more connections across and between resources, and enhance consistency of bibliographic data. In addition to

bibliographic systems, automatic indexing is used today in a wide variety of applications such as e-mail filtering and focused crawling (see Sebastiani, 2002, p. 6-9).

Software vendors and experimental researchers speak of the high potential of automatic indexing tools. While some claim to entirely replace manual indexing in certain subject areas (e.g., Purpura & Hillard, 2006; Roitblat, Kershaw, & Oot, 2010), others recognize the need for both manual (human) and computer-assisted indexing, each with its advantages and disadvantages (Anderson & Perez-Carballo, 2001; Hagedorn, 2001; Lykke & Eslau, 2010; Svarre & Lykke, 2014). Reported examples of operational information systems include NASA's machine-aided indexing which was shown to increase production and improve indexing quality, although there were other contributing variables (Silvester, 1997); and the *Medical Text Indexer* at the US National Library of Medicine (2010), which by 2008 was consulted by indexers in about 40% of indexing throughput (Ruiz, Aronson, & Hlava, 2008).

Hard evidence on the success of automatic indexing tools in operating information environments is scarce; research is usually conducted in laboratory conditions, excluding the complexities of real-life systems and situations. However, results of evaluating subject indexing *per se* and out of context may be meaningless. For example, having reviewed a large number of automatic indexing studies, Lancaster concluded that the research comparing automatic versus manual indexing is seriously flawed (2003, p. 334).

Our purpose is to develop a framework for the evaluation of subject indexing or classification quality. We base the framework on a comprehensive overview of methods for evaluating subject indexing quality with emphasis on evaluation of automatic indexing tools. The aim is to develop a framework that will offer insight into whether to use an existing tool or how to improve such tools. The primary scenario we have in mind is retrieval in a circumscribed collection using one or more of the following approaches to retrieval: Boolean retrieval; a subject directory structure through which the user can navigate; constraining search results through facets or filters from which the user can select specific criteria.

The remainder of the paper discusses the background and rationale and then presents the framework. The conclusion sums up and discusses implications and further research. This paper focuses on methods of data collection and analysis. A companion (Soergel & Golub, 2015) reviews and analyzes measures of indexing performance and retrieval performance.

Background

Terminology

For simplicity, we use the term *indexing* broadly to refer to any form of assigning subjects – uncontrolled terms or controlled vocabulary terms or classes from a classification scheme – to denote a document's (or other entity's) relevance to a topic, including aboutness. *Subject indexing*, narrowly defined, is commonly used for the assignment of multiple typically elemental or moderately precombined subject terms (with a typical range from 3 – 20) and *subject classification*

for the assignment of a few, typically one, highly precombined *subject class(es)* from a subject classification scheme to express the major theme of the document. The purpose of subject indexing is to allow retrieval of a document from many different perspectives. The main purpose of subject classification is to group similar documents together to allow browsing. *Subject assignment* might be a better term for the inclusive concept, but for the sake of readability we use *subject indexing* or just *indexing*. The term *subject metadata generation* is also used for subject assignment, and *text categorization* for subject classification, where *subject* is broadly construed to include, for example, document genre. In the context of automatic methods, *automatic classification* or *automatic (text) categorization* may refer to the assignment of a class or category from a pre-existing scheme (the meaning we use in this paper) or it may mean to discover a scheme suitable for the collection at hand and simultaneously assign to a document one (or more) of the classes discovered. This may be achieved, for example, through clustering.

Relevance

The purpose of indexing is to make relevant documents retrievable. A reminder on the complex nature of relevance, and therefore the complex nature of indexing and its evaluation, is thus in order. There are many possible relationships between a document and a query, making relevance complex. Relevance is often determined through human judgment, introducing an element of subjectivity. Borlund (2003) emphasizes that relevance is *multidimensional* and *dynamic* – users use many different criteria in assessing relevance and perception of relevance can change over time for the same user. There are also different classes, types, degrees, and levels of relevance. A critical review of the notion of relevance in terms of its nature, manifestations, related theories and models is given by Saracevic (2007a, 2007b). Huang and Soergel (2013) build on this review and provide a new conceptual framework for relevance, which comprises relevance-as-is and relevance-as-determined. Relevance-as-determined is the result of the assessment or determination of relevance-as-is by a person or a computer system based on representations of the information object and the information need. This determination can be made before, during, or after use of the information. Relevance-as-determined is always an imperfect approximation, influenced by the given conditions and by the limited knowledge and processing capacity of the determining agent (expert, user, or computer system). Topical relevance is of central importance and goes beyond simple topic or term matching but involves over 200 fine-grained topical relevance relationships.

Earlier, Soergel (1994) quotes three relevance types as central to good subject indexing and to assessing retrieval performance: topical relevance, pertinence, and utility. He defines topical relevance as a relationship between a document and a topic, question, function, or task; a document is topically relevant for a question if it can "shed light on the question". A document is pertinent if it is topically relevant and if it is appropriate for the person, that is, if the person can understand the document and apply the information gained. Finally, a document has utility if it is pertinent and makes a useful contribution beyond what the user knew already; a pertinent document lacks utility; if the user is already familiar with its content.

In spite of the dynamic and multidimensional nature of relevance, in practice evaluation of information retrieval systems is often reduced to using pre-existing relevance assessments. The influential Cranfield tests (Cleverdon, Mills, & Keen, 1968) have set up the prevalent methodology for information retrieval evaluation which involves a gold standard: a test collection consisting of a set of documents, a set of 'topics', and a set of relevance assessments (Buckley & Voorhees, 2000). A 'topic' is a description of the information being sought. Relevance assessments specify the documents that should be retrieved in response to each topic; they are usually binary, making no provisions for the different degrees of relevance. It is possible to tell whether A is relevant to B, but specifying precisely in what way A is relevant to B and figuring out how a piece of information fits into the overall structure of a topic and how it contributes to the user's thinking and reasoning about the topic is much harder. Thus Huang and Soergel (2013) argue for a shift in emphasis from an entity-focused approach to the relationship-focused approach to conceptualizing relevance which they propose. They emphasize a high priority for research into different types of relevance relationships for fully revealing the concept's depth and richness (Huang and Soergel, 2013, p. 31).

Indexing characteristics

The ISO indexing standard (*ISO 5963:1985*, confirmed in 2008 (International Organization for Standardization, 1985)) gives a document-oriented definition of manual subject indexing as a process involving three steps:

1) determining the subject content of a document,
2) a conceptual analysis to decide which aspects of the content should be represented,
3) translation of those concepts or aspects into a controlled vocabulary.

In contrast, request-oriented (problem-oriented, user-oriented) indexing (Brenner & Mooers, 1958; Soergel, 1985; Fidel, 1994) places the focus on the potential users and uses; the indexer's task is to understand the document and then anticipate for what topics or uses this document would be relevant.

The first and second steps in the ISO definition are related to the concept of aboutness, the document's subject matter or topics. Aboutness and the complexity of determining appropriate subject terms has been widely discussed (for an overview see Lancaster, 2003, pp. 13-19). Aboutness is not always explicitly stated; determining aboutness may involve knowledge that is shared by the creator(s) and user(s) of the text. This in turn is influenced by many cognitive factors, such as interest, task, purpose, knowledge, norms, opinions and attitudes, which then determine the meaning the text has for its user (Moens, 2000, p. 12). Request-oriented indexing considers not only aboutness but also anticipated relevance of a document to topics, purposes, tasks, and problems to be solved, introducing further context-dependency and diversity. It may not be possible for any single indexer to determine all ideas and meanings that might be associated with a document since there always might be potential ideas and meanings which different people at different times and places might find in the document (for example, Mai, 2001, p. 606), but indexers need to do the best they can, being cognizant of the needs of the intended users. Social tagging offers the opportunity of indexing a document from a larger number of perspectives.

The second and third steps in the ISO definition are based on a specific policy with respect to the document collection and target user groups, for example in terms of exhaustivity (the importance threshold applied in selecting concepts for indexing, with a low threshold resulting in a larger number of concepts indexed, high exhaustivity) and specificity (the hierarchical level at which to index), or whether target user groups are school children, laymen, or specialists etc. Thus, evaluation of automatically assigned against manually assigned subjects must consider the collection's indexing policies. A subject correctly assigned in a high-exhaustivity system may be erroneous in a low-exhaustivity system (Soergel, 1994). Lancaster (2003, pp. 86-87) lists indexing errors as:

- errors related to the exhaustivity policy (too many or too few subjects become assigned),
- errors related to specificity (usually: the assigned subject is not the most specific one available),
- errors of omitting important subjects,
- errors of assigning obviously incorrect subjects.

Related to this is the question of indexing consistency between different indexers (inter-indexer consistency) or between different sessions by the same indexer (intra-indexer consistency) (for an overview see Lancaster 2003, pp. 68-82). Consistency can be measured at the level of concepts versus at the level of terms used to express the concepts. Different people, whether end users or professional indexers, assign different subjects – different concepts or different terms for the same concepts – to a document. Markey (1984) reviewed 57 indexer consistency studies and reported that consistency levels ranged from 4% to 84%, with only 18 studies showing over 50% consistency. There are two main factors that seem to affect it:

1) Higher intended exhaustivity and specificity of subject indexing both lead to lower consistency, i.e. indexers choose the same first index term for the major subject of the document, but the consistency decreases as they choose more subjects;
2) The bigger the vocabulary (the more choices the indexers have), the lower the probability that they will choose the same terms (Olson & Boll, 2001, pp. 99-101).

Lancaster (2003, p. 71) adds further possible factors: subject indexing based on uncontrolled versus controlled vocabulary terms, specificity of the vocabulary, characteristics of subject matter and its terminology, indexer factors (e.g., experience in training), tools available to the indexers, and length of item to be indexed.

Furthermore, indexing can be consistently wrong as well as consistently good (Cooper, 1969). This means that high indexing consistency is not necessarily a sign of high indexing quality (Lancaster, 2003, p. 91). High consistency is a necessary but not sufficient condition for high correctness and thus indexing consistency should not be used as the prime indicator of indexing correctness (Soergel, 1994). On the other hand, high indexing correctness results in high consistency. Rolling (1981) also shows how indexing consistency, indexing quality, and indexing effectiveness do not necessarily evolve proportionally. He defines indexing quality as whether the information content of an indexed document is

accurately represented (a document-oriented definition), and indexing effectiveness as whether an indexed document is correctly retrieved every time it is relevant to a query (this brings in request orientation).

Nevertheless, indexing consistency has often been used as a measure of the quality of a gold standard and as a measure of indexing quality in operational systems without reference to a gold standard, as suggested by Rolling (1981). Thus, for example, Medelyan and Witten (2006) evaluated a thesaurus-based indexing algorithm using the common precision and recall measures, but also proposed defining the "gold standard" in indexing as the level of inter-indexer consistency, their aim then being to develop an automatic indexing method that is as consistent with a group of indexers as the indexers are among each other.

Terms assigned automatically but not manually might be wrong or they might be right but missed by manual indexing. Also, subject indexing involves determining subject terms or classes under which a document should be found and what it could be used for; this goes beyond simply capturing what the document is about and is generally done better by people than by computer programs. However, automatic indexing algorithms that use rules learned from a good training set might find such terms, but human indexers who are not well trained might miss them. For example, Roberts and Souter (2000), Ribeiro-Neto et al. (2001) and Golub (2006b) report how automated tools assigned many index terms that should not have been assigned, but that they also added on average one index term per document that was missed by human indexers.

Evaluation in automatic subject indexing is mostly conducted under controlled conditions, ignoring issues such as exhaustivity and specificity policies at hand, document-oriented versus request-oriented correctness, inter- and intra-indexer consistency. As Sebastiani (2002, p. 32) puts it, "…the evaluation of document classifiers is typically conducted experimentally, rather than analytically. The reason is that… we would need a formal specification of the problem that the system is trying to solve (e.g., with respect to what correctness and completeness are defined), and the central notion… that of membership of a document in a category is, due to its subjective character, inherently nonformalizable."

The indexing evaluation framework. Overview

The framework includes three complementary approaches and allows for triangulation of methods and exploration of multiple perspectives and contexts:

1. Evaluating indexing quality directly through assessment by an evaluator or by comparison with a gold standard.
2. Evaluating indexing quality in the context of an indexing workflow.
3. Evaluating indexing quality indirectly through retrieval performance.

The framework emphasizes evaluation in real-life environments or scenarios that simulate real-life environments closely.

The atomic unit of data in evaluating indexing is the correctness or goodness of an assigned index term. If all good index terms are known, the quality of indexing of a document by a given method can be computed as a function of

good index terms assigned, bad index terms assigned, and good index terms missed. In Approach 1 experts assess index terms assigned by the method to be evaluated or they provide assessments to be "frozen" in a gold standard collection. In Approach 2, an indexer assesses the terms suggested by a computer-assisted indexing system. In Approach 3, the contribution of an index term to correct and incorrect retrieval across multiple queries determines the goodness of the term. The approaches can be combined: Approach 2 can be followed by using Approach 1 to evaluate both the output of the computer-assisted indexing system and the final output produced by the indexer. Approach 3 has an additional measure, retrieval performance achieved by different methods of assigning index terms. This could be used as the sole measure, or it could be used in conjunction with Approach 1 to see whether a system that gets better indexing quality as determined by direct measurement also gets better retrieval performance.

We discuss each of the three approaches in turn. First we review important literature relevant to the approach. Then we put ourselves into the shoes of someone tasked with evaluating indexing quality with limited resources (for example, as the basis for deciding on an automatic indexing tool) and make recommendations for such an evaluation. Our recommendations are based on the literature review and on our collective experience. We are well aware that these recommendations represent only our best intuition and that they will not deliver unassailable quantitative assessments, but we do believe that they will lead to insights that allow for improvements in automatic indexing and more informed decisions.

Evaluating indexing quality directly through assessment by an evaluator or through comparison with a gold standard

Two main approaches to direct evaluation of indexing quality exist: (1) ask evaluators to assess index terms assigned, and (2) compare to a gold standard.

Evaluation through direct assessment by an evaluator

Rosenberg (1971) compared two subject indexing methods. He used a panel of judges to rate assigned index terms both for their appropriateness to the document and their usefulness as access points. Golub & Lykke (2009) evaluated automatic assignment of subject classes from the Engineering Index classification to a collection of 19,000 Web pages for use in hierarchical subject browsing. Forty participants performed four controlled search tasks developed following the methodology of Borlund (2003), using simulated work tasks which represent situational relevance and thus allow for testing subjective relevance in a controlled environment. Once users identified the relevant class for the search, they assessed for each of the top 10 documents found the class assigned, using a three-point scale (correct, partly correct and incorrect). There were large differences among participants in their assessments – for a number of web pages the class assigned was assessed as correct, partly correct, and incorrect by different participants. The differences in assessment

illustrate the subjective nature of assessments in spite of detailed instructions. Tsai, McGarry, & Tait (2006) in a qualitative evaluation of automatically assigned keywords for images combined two methods for complete evaluation and deeper understanding: (1), they had people evaluate automatically assigned subject index terms and (2) they had people assign subject terms to be used as the gold standard.

Evaluating indexing quality through comparison with a gold standard

A gold standard is a collection in which each document is assigned a set of subjects that is assumed to be complete and correct. 'Complete' means that all subjects that should be assigned to a document are in fact assigned, and 'correct' means that there are no subjects assigned that are irrelevant to the content. Once a gold standard is developed, the results of any method of subject assignment (manual or automatic) applied to the collection at hand can be evaluated by comparison with the gold standard. For any document indexed by some method, one can then determine the number of correctly assigned terms, the number of incorrectly assigned terms, and the number of missed terms and compute, for example, precision and recall defined for index term assignment.

In creating a gold standard, information experts, subject experts, or potential or real end users assign presumably correct index terms. However, due to the subjective element of assigning subjects to documents, a reliable gold standard should involve a group of experts reaching consensus decisions, considering all relevant aspects from which a document should be described and which subjects would be most appropriate. This group would include information experts, subject experts, and end users. Note that the goal of indexer consensus is based on the assumption that for every document there is one best set of index terms for all purposes and users. This is clearly not the case. One might use a group of experts to assign subjects from multiple perspectives providing diversity. An information retrieval system should allow for such diversity catering to different users, introducing even more complexity for the evaluation of indexing. In this paper we do not consider this complexity.

Because of these problems, any gold standard must itself be evaluated. Evaluation of a gold standard is concerned with measuring the quality of indexing, this is often measured by having several indexers index the same set of documents and then measuring the agreement of subjects assigned to each document (document-based indexing consistency). This is in spite of the fact that consistent indexing can be consistently bad indexing (see section *Indexing characteristics* on indexing consistency and indexing quality).

Even with all these precautions, basing evaluation only on a gold standard presents problems that have been discussed in the literature (Soergel, 1985, chapter 18). The validity and reliability of results derived solely from a gold-standard evaluation remains, to our knowledge, unexamined. Another downside of using a gold standard is that a set of subject index terms cannot be judged as 'correct' or 'incorrect' in any absolute sense, as there is no one 'best' set of terms. Such a claim would imply knowing all the requests that will be put to the information retrieval system (Lancaster, 2003, p. 86). There is an alternative to relying solely on an a priori, gold standard: Assess the terms not in the overlap of E, the

indexing method to be evaluated, and G, the gold standard (terms assigned by E but not G and terms assigned by G but not E), thereby improving the gold standard. In this approach, the terms assigned by both E and G are assumed to be correct, but they could be checked again.

The gold standard approach is commonly used in the text categorization community, an important community in automatic subject indexing and classification (for an overview of text categorization see Sebastiani, 2002; for a discussion of approaches to automatic subject classification see Golub, 2006a). In text categorization, a correctly indexed collection is commonly used both for training the algorithm and for evaluation: The algorithm first 'learns a classifier from a sub-group of documents (the training set). The classifier is then applied to the documents in the remaining collection (the test collection or gold standard).

Several collections are available for testing text categorization, mostly as applied to news stories:

- The Reuters collection, a set of newswire stories classified under categories related to economics.
- The Reuters Corpus Volume (Lewis et al., 2004), a more current and better-documented collection of 800,000 manually categorized newswire stories with about a hundred categories. It has been prepared for text categorization experiments, with errors and inconsistencies of the original version removed.
- 20 Newsgroups DataSet (1998), 20 newsgroups where each newsgroup corresponds to one category.
- Many studies have used ad-hoc document collections. For example, Mladenic (1998) used Yahoo directory categories and corresponding documents as a test collection. Chung, Pottenger, & Schatz (1998) used 76 Compendex documents in the field of programming languages as a test collection. Each document had assigned Compendex controlled terms, uncontrolled terms, automatic terms, as well as terms assigned by evaluators. Each term was judged as highly relevant, relevant, neutral, or irrelevant, and term precision and term recall were calculated for each of the four types of subject index terms.

Sebastiani (2002), discusses evaluation in text categorization; he states that evaluation results are influenced by experimental setup and lists the following factors which must be the same in order to allow for comparison across different studies: exactly the same collection (i.e., same documents and same assigned categories), the same 'split' between training set and test set, and the same evaluation measure and its parameter values. His general assessment is that "… although the TC [text categorization] community is making consistent efforts at standardizing experimentation protocols, we are still far from universal agreement on evaluation issues and, as a consequence, from understanding precisely the relative merits of the various methods." (p. 35). Lewis et al. (2004, p. 362) wrote that existing text categorization collections suffer from one or more of the following weaknesses: "few documents, lack of the full document text, inconsistent or incomplete category assignments, peculiar textual properties, and/or limited availability". Moreover, documentation on how the collections were created and on the nature of their category systems is often missing. They also state how, even if the gold standard

collections were perfect, there would be a continuous need for new ones because algorithms can be tuned to perform well on a dataset; in order for algorithms to advance, they need to be tested on new collections.

Some studies in in automatic subject indexing have used existing library catalog data as the gold standard. Examples include Plaunt and Norgard (1997) using the INSPEC thesaurus in engineering, Aronson et al. (2004) using MeSH, and Paynter (2005) using the Library of Congress Subject Headings (LCSH). Terms assigned through social tagging can also be in the evaluation of automatic indexing. However, as discussed in Section *Indexing and relevance characteristics,* the manual indexing is often conducted without thorough consideration and quality control and not vetted through a consensus and consistency among subject indexers may vary considerably, from 4% to 84 % (Markey, 1984).

Recommended approach for evaluating indexing quality directly

We recommend the following steps:

- select 3 distinct subject areas that are well-covered by the document collection;
- for each subject area, select 20 documents at random, possibly stratified by type of document;
- 2 professional subject indexers assign index terms as they usually do (or use index terms that already exist);
- 2 subject experts assign index terms;
- 2 end users who are not subject experts assign index terms;
- assign index terms using all indexing methods to be evaluated (for example, several automatic indexing systems to be evaluated and compared);
- prepare document records that include all index terms assigned by any method in one integrated listing;
- 2 senior professional subject indexers and preferably 2 end users examine all index terms, remove terms assigned erroneously, and add terms missed by all previous processes.

Decisions on the number of indexers, evaluated documents, and types of measures applied in order to make the evaluation reliable and repeatable must consider the context and the resources available. To our knowledge, there are no studies that shed light on how these numbers affect the quality of the results. Statistical power calculations might be useful but problematic since nothing about underlying distributions is known. It stands to reason that higher numbers, especially of documents, lead to more accurate results. In most studies the question is what one can afford. Intuitively, less than 20 documents per subject area would make the results quite susceptible to random variation. So there is a point where what one can afford is not worth doing because the results could not be relied upon. As a further complication consider that some index terms are more important for the anticipated uses of a retrieval system than others. Looking at the results of a study done with small numbers one might be able to ascertain whether collecting more data is necessary (a very informal application of the idea of sequential tests).

Evaluating tools for computer-assisted indexing directly in the context of an indexing workflow

Automatic indexing tools can be used for computer-assisted (computer-aided, machine-aided) indexing, and some tools are designed for that purpose; such tools form part of a document processing workflow. Evaluating the quality of computer-assisted indexing tools should assess the value of providing human indexers with automatically generated index term suggestions.

Several studies have been conducted as part of the US National Library of Medicine's Indexing Initiative (2009, 2010) testing the *Medical Text Indexer* (MTI). Aronson et al. (2004) conducted an experiment with ten volunteer indexers who each indexed the articles from one journal issue, selecting from about 25 MTI-suggested terms and adding their own. The MTI module was integrated into an experimental version of the regular indexing software. For each article, the indexer completed a questionnaire about the MTI-suggested terms. After indexing all articles the indexers completed a final questionnaire about the general experience with MTI. A later study (Ruiz & Aronson, 2007) investigated human factors affecting adoption of MTI after it was integrated into the regular indexing workflow. The study covered evaluation of tasks performed (which ones are more time-consuming and intellectually challenging), usage of tools available in the document creation and maintenance system (of which MTI is one), and usage of MTI (self-reported). The study focused on *perceived usefulness,* or the degree to which the person believes using a system would enhance his/her job performance, and *perceived ease of use*, or the degree to which a person believes that using a system would be free from effort (these factors are taken from a general model of the acceptance and use of new technologies proposed by Davis, Bagozzi, & Warshaw, 1989).

Tonkin and Mueller (2008) studied time saved through computer-assisted indexing as suggested in Polfreman and Grace (2008). They used two groups of professional subject indexers to index a test set of documents. The first group indexed the first half of the documents without automatic term suggestions and the second half with, and the second group did the same in reverse order.

Bainbridge and Witten introduced a number of tools to support the librarian's workflow (see, for example, Bainbridge & Witten, 2008). However, they did not evaluate these tools from the perspective of human-computer interaction.

Only a few measures for evaluating computer-assisted indexing tools have been used or proposed. Silvester (1997) used the following:

1. Match rate, proportion of automatically suggested terms that the indexer selects to use (precision);
2. Capture rate, proportion of indexer-assigned terms present in automatically suggested terms (recall);
3. Consistency factor, overlap of the two term lists.

See also the overview of metrics used for evaluating automatic metadata by Polfreman and Grace (2008), where *time saved* is suggested. One should also look at any change in the *quality of indexing*.

Recommended approach for evaluating tools for computer-assisted indexing directly in the context of an indexing workflow

A study of an existing tool that is considered for adoption or a demonstrator of a tool under development should be carried out in four phases:

1) Collecting baseline data on unassisted manual indexing.
2) A familiarization tutorial for indexers, concluded with a short summative interview.
3) An extended in-use study. Observe practicing subject indexers in different subject areas. Determine the indexers' assessments of the quality of the automatically generated subject term suggestions, identify usability issues, and evaluate the impact of term suggestions on terms selected.
4) A summative semi-structured interview.

The result will be an understanding of the effects of computer-assisted indexing in practice.

Such an evaluation should consider

- the quality of the tool's suggestions,
- the usability of the tool in the indexing workflow (Blandford et al., 2004);
- the usability of the system being studied;
- the indexers' understanding of their task;
- The indexers' experience with computer-assisted indexing tools
- the users' structuring and organization of the performance of the task;
- cognitive support the system provides to users in the course of their decision-making;
- the resulting quality of the final indexing
- time saved.

Experience of the indexers with the types of tools to be evaluated is a prerequisite for obtaining realistic and reliable data on the practical value of such a tool, but few indexers have such experience. A demonstrator prototype would allow indexers to gain such experience. If the period of familiarization is too brief, the perceived usability may be lower than the potential usability once a good working understanding of the tool is achieved. The familiarization period itself will uncover how easy it is to *learn* to use the system. Some data should be collected over a longer period to observe how the use of the tool changes as indexers' familiarity with it increases.

Evaluating indexing quality indirectly through retrieval performance

Since the major purpose of subject indexing is successful information retrieval, the quality of subject indexing can and should be studied in that context, preferably involving real user requests and real end users (see Lancaster, 2003, p. 99). This section discusses assessing indexing quality by comparing retrieval results from the same collection using indexing from different sources with emphasis on detailed analysis of how indexing contributes to retrieval successes or failures. It reviews issues of retrieval testing of particular importance in this context, including the effect of search tasks and building test collections (by necessity this review is rather eclectic).

Soergel (1994) gives a logical analysis of the effects of subject indexing on retrieval performance. He identified indexing devices and viewpoint-based and importance-based indexing exhaustivity, specificity, correctness, and consistency as affecting retrieval. He concludes that retrieval performance depends chiefly on the match between indexing and the requirements of the individual query and on the adaptation of the query formulation to the characteristics of the retrieval system. This complexity must be considered in the design and testing of retrieval systems: "…indexing characteristics and their effects on retrieval are so complex that they largely defy study in artificial test situations. Most experiments fail to account for important interactions among factors as they occur in the real world, and thus give results that mislead more than they enlighten, results that have little meaning for the assessment or improvement of operational retrieval systems" (p. 589).

In typical information retrieval tests, retrieval effectiveness is established by comparing documents retrieved by an algorithm to documents assessed as relevant by assessors. The set of assessments on which documents are relevant to which query is often used as the gold standard. As with gold standards for subject indexing, gold standards for relevance assessments are subjective and vary depending on a number of factors. Like subject index terms, relevance assessments are known to be inconsistent among assessors as well as for the same assessor at different times (see section *Indexing and relevance characteristics* above on the concept of relevance).

Saracevic (2008) analyzed seven studies that examined the effect of inconsistent relevance assessments on the ranking of competing retrieval systems (rather than measuring absolute performance); he concluded that while some claim that inconsistent relevance assessments have no effect, others indicate that effects do occur and that more research is needed to explore this further. He also pointed out a powerful influence of an early retrieval study by Gull in 1956 on selection of method for obtaining relevance assessments. Gull's study reported that two groups of assessors could not agree on relevance assessments and Saracevic claims that it has since become common practice not to use more than a single assessor for establishing a gold standard (p. 774), effectively ignoring the problem of inconsistency.

Because of the above factors, it is important to

(1) build the gold standard of relevance assessments carefully;

(2) evaluate indexing – manual or automatic – in the context of end-user information retrieval.

Review of evaluation studies that focus on the effect of indexing on retrieval performance

Real-life testing is very hard. This section discusses studies and methods that address this issue either by feasible real-life procedures or substitutes. These methods can be used to analyze indexing by a single system or to compare the performance of two or more systems, manual or automatic.

Lancaster's (1968) classic MEDLARS study is one of very few studies that involved real users, real queries, and actual results given to users; it analyzed retrieval for 300 MEDLARS requests, including effects of indexing and failure analysis (Saracevic, 1998). Hliaoutakis, Zervanou, & Petrakis (2009) examined the effects of two automatic indexing methods on retrieval of abstracts and full-text documents. Evaluation for abstracts was based on 64 TREC queries and evaluation for full-text documents on 15 TREC queries; both used TREC relevance assessments.

Lancaster (2003, p. 87) suggested a simulation method as follows:

1) Select a group of documents (exact number not given).
2) For each document, compose, e.g., three questions for which it can be considered an important response (one question based on the central theme of the document and two on secondary but still important themes). (Note: This method is problematic if the questions are not representative of real user questions).
3) Have experienced search analysts construct a query formulation for each question.
4) Independently, have the documents indexed in the normal way (or use existing indexing).
5) Compare the indexing to the query formulations to determine whether or not the relevant items are retrievable on the terms assigned. (One should also check whether other relevant documents were found).

This method was first used by Albright (1979) in a study of the retrievability of items in an information retrieval system. He used 10 simulated questions/topics and selected for each 80 relevant documents from Index Medicus. He then used several queries for each topic until the queries retrieved among them all 80 documents known to be relevant to the topic. On average 44 different terms had to be used to retrieve all items known to be relevant to a given topic. This would require the user to be very knowledgeable and persistent in order to achieve high recall.

Lykke and Eslau (2010) compared retrieval performance of automated categorization and free-text indexing in the retrieval system of a pharmaceutical company. They selected ten genuine topics from the company's search log and conducted three searches for each search topic:

1) a search on controlled terms assigned manually from the company's domain-specific thesaurus;
2) free-text search with terms taken from the real-life question;
3) free-text search with query expansion using the company thesaurus.

The queries were based on the terms that the original searcher in real life used to formulate the first query they submitted in an interactive search for one topic. In the test, only one query was searched for each topic, there was no interaction.

Retrieved documents were assessed for relevance on a 4-point scale by the original searcher according to the work task situation. Results, averaged over the ten topics, are shown in the following table:

Table 1. Relative precision and recall results for the three types of searches

	1 Controlled-terms search	2 Free-text search	3 Free-text w/ expansion
Relative recall	24%	41%	89%
Precision	17%	33%	24%

The findings concerning precision are remarkable, as human indexers should be better at weighing the significance of document subjects, especially in the context of enterprise retrieval, where retrieval is embedded in and targeted to specific information tasks. It could be that the particular search topics could be better and more specifically expressed natural language. Since the test used single-query searches without interaction, the results do not necessarily apply to interactive information retrieval.

Svarre and Lykke (2014) compared retrieval performance using (1) automated categorization and (2) full text indexing in the government intranet used by the Danish tax authorities. Thirty-two participants performed each 4 search topics (sessions), 3 simulated and 1 real-life search topics. In all, 128 sessions were carried out, 64 in each of the test systems, with a total of 564 queries. Searching behaviour and search outcome was documented by search logs, relevance assessments on a scale from 1 to 3, and post-search interviews. In addition, the post-search interviews produced qualitative data - insight into the searchers' considerations and decisions when they respectively browsed the taxonomy or made keyword searching. Search performance was measured by two metrics: (a) query success: percentage of queries retrieving at least one document with relevance score 2 or 3 and (b) session success: percentage of sessions that solved the work task problem. The full-text indexing performed best, 31% in query success compared to 22% for the automated categorization, and 89% in session success compared to 84%. Different causes were found for the result. The more restrictive AND operator was used with the same frequency in both systems, resulting in very small result sets in the system based on automated categorization. Further, some participants expressed trouble finding suitable categories in the categorization to match their queries due to lack of knowledge of the taxonomy.

Effect of search tasks and search requests

Work tasks and resulting search tasks influence both search strategies and relevance assessments. For example, results of a search for depression would be assessed differently by a physician who wants to brush up on the latest developments in depression and by a medical researcher preparing a thorough review of all aspects of depression. Therefore, retrieval studies must consider work tasks and search tasks (Liu & Belkin, 2015).

Ingwersen and Järvelin (2005, p. 327) classify work task and search tasks scenarios on a natural to artificial continuum:

1) Natural manifestations carried out in real life, where the problem for research is how to control the variables involved;
2) Simulated situations designed for research and involving a specified but invented scenario (such as suggested by Borlund, 2003), where the scenario acts as a controlled variable of context; and,
3) Assigned information requests meant to represent information needs (TREC 'topics'), which do not operate with a work task but are highly controlled and artificial.

Work tasks can also be classified by complexity; see, for example, Kim and Soergel (2005), Belkin et al. (2014). Ingwersen & Järvelin (2005, p. 337) classify search requests along six dimensions:

1) Well defined versus vaguely defined;
2) Generic versus specific (e.g., specific would be referring to a particular person or event);
3) Simple versus complex (the latter involving complex relationships between several concepts or facets);
4) Short versus wordy;
5) Real versus assigned information needs [authors' note: this does not fit here];
6) Articulated versus implicit needs.

We give some examples of studies of the effects of these variables. Iivonen (1995) used four different types of search requests. Each request was formulated by several [exact number not given] searchers. Inter- and intra-searcher consistency differed significantly by search type. Kim (2008) demonstrated considerable effects of specific versus general types of tasks on search behavior.

Beaulieu (2003) describes two user information retrieval studies, each using different types of tasks. Study 1 used a fact-finding task and a research-based searching task, and Study 2 a simple and complex searching task. Suomela and Kekäläinen (2006) used a general simulated work task, giving rise to four detailed search tasks. The work task was writing a fairly large report on a given theme. Golub and Lykke (2009) investigated whether it is meaningful to use the Engineering Index classification scheme for browsing, and then, if proven useful, to investigate the performance of an automated classification algorithm based on the classification scheme. A user study was conducted in which users solved four controlled searching tasks in which they (1) browsed the classification scheme in order to examine the suitability of the classification systems for browsing and (2) judged the correctness of the automatically assigned classes. Each task was from a different subject field and of a different character: two in the basic sciences as applied in engineering, two in general engineering; two at the fifth hierarchical level of the classification scheme, and two at the fourth hierarchical level; in two search tasks topic names were the same as class captions and in the other two topic names and class names were entirely different.

Test collections in information retrieval

Information retrieval tests. commonly use gold-standard test collections consisting of documents, search requests (often called topics), and relevance assessments (Buckley & Voorhees, 2000). Ideally, relevance assessments are made for every document-request pair (as was done in the Cranfield experiments). However, with realistic-size test collections this is prohibitively expensive, so different ways for preselecting documents to be assessed for a given request have been used (see examples below). Often the relevance assessments are not revisited once made. However, it is much safer to assess documents retrieved by a system again because this may uncover documents whose relevance was previously missed (Soergel, 1985, p. 389). The TREC Legal Track has a process for such a post-retrieval check of relevance assessments; many assessments are changed from non-relevant to relevant.

Sparck Jones and van Rijsbergen (1976) give a set of recommendations for an ideal test collection that would make results statistically valid, adequately control the variables, and make results transferable to real systems, but for qualitative studies these do not apply.

1) Number of documents: < 500 is of no real value, 1,000-2,000 is acceptable for some purposes, and > 10,000 is needed in some cases.
2) Number of search requests: < 75 are of little use, 250 are usually acceptable, 1000 are sometimes needed.
3) The ideal collection should be controlled but also exhibit variety in content, type, source, origin, date, natural language, document and request length. (Note: This does not apply if one wants to study a specific type of document or request.)
4) While the ideal for relevance assessments is exhaustive judging, this may simply not be feasible; thus, parallel searches should be conducted to establish recall estimates.
5) Graded relevance assessments, from not at all to marginal to high.

Today the best known information retrieval collections exist within the TREC initiative (Text REtrieval Conference, 2010). There are a number of different collections organized around tasks called tracks; for example, the ad hoc track (the task is accurate ranking to a given topic), the interactive track, Web, video, cross-language, speech and others (Voorhees & Harman, 2005). TREC uses subject experts to come up with 50 topics for which there is a reasonable number of relevant documents in the test collection. Participating teams use their systems to find documents for each topic and the top 100 documents from each team are combined into a pool. Relevance is assessed for each document in the pool. Documents not in the pool are simply considered not relevant. The limits of judging the rest as non-relevant have been explored by Buckley et al., 2007); one pool consisted only of documents with titles containing topic words because this was the ranking principle that all the participating algorithms used; however, these documents then squeezed out other

kinds of relevant documents. (This point was already made by Soergel (1985, p. 391) in a critique of the Cranfield retrieval tests). Urbano, Marrero, & Martín (2013) evaluated over 40 TREC collections and found that they are not as reliable as generally accepted and that 50 queries may not be enough.

Test collections are also available through INEX, INitiative for the Evaluation of XML retrieval, set up in 2002 (Lalmas & Tombros, 2007). INEX uses a four-graded scale of relevance: highly relevant, fairly relevant, marginally relevant, and non-relevant.

Still another source for test collections is CLEF (Cross-Language Evaluation Forum), (Braschler & Peters, 2004). CLEF includes a speech retrieval test collection consisting of 8,100 speech documents (segments of oral history interviews from the Shoah Foundation collection) and 70 topics developed from actual user requests (Oard et al., 2004). Eight history and information studies students, who were thoroughly trained and met regularly, served as assessors. They conducted very thorough searches to identify potentially relevant documents and then assessed relevance along four different types of (perspectives on) relevance (direct, indirect, context, comparison) on a five-point scale. They would often cycle back to expand the search. From these detailed data (that are still available) a simplified test collection was generated. In conformance with TREC definitions, direct and indirect relevance assessments were considered for determining what is "relevant" and the graded assessments were collapsed to binary values.

Relevance assessments in building retrieval test collections

Relevance assessors can be recruited from different groups:

- end users,
- subject experts (one or consensus of a group),
- information professionals,
- 'bystanders' belonging to none of the above – for example, students asked to do a given task of assessment (Saracevic, 2008),
- crowdsourcing – engaging a large group of unknown people (a large number of 'bystanders') to acquire relevance assessments (see, for example, Hosseini et al. 2012; Kazai, Kamps, & Milic-Frayling, 2013; Venanzi et al. 2014).

In most cases it is not real users who are the ultimate relevance assessors for information retrieval test collections.

Because of the subjective nature of relevance (see ***Error! Reference source not found.***), it has been argued that ideally each topic would be assessed by at least three assessors and that a consensus process would provide maximum consistency (Hripcsak & Wilcox, 2002; Colosimo et al., 2005). Having assessors from more than one of the groups listed above would be especially beneficial. However, the general assumption of "one size fits all" relevance assessments is clearly not realistic; the field needs to move to evaluating system performance for different types of users. While crowdsourcing has been used to address the issue of scalability, in particular when comparing search engines, research has

shown many complex factors contributing to varying quality of the resulting relevance assessments, including: variables such as the level of pay offered, the effort required to complete a task and the qualifications required of the workers; factors like motivation, interest, familiarity with the topic, perceived task difficulty, and satisfaction with the offered pay; and interactions of those factors with aspects of the task design (Kazai, Kamps, & Milic-Frayling, 2013). Thus crowdsourcing of relevance assessments does not fit well with an evaluation approach that is situated in the context of real-life systems and situations and takes into account the complexities encountered in such systems.

The perspective(s) used in judging relevance are important. For example, TREC uses a narrow perspective: To be relevant, documents must be about the topic (direct or indirect/circumstantial relevance). A document that might be useful to a user by providing context, or allowing comparison with a different situation, or that might be useful in other ways for dealing with the topic is not considered relevant in TREC. (See *Relevance* and the brief description of the CLEF speech retrieval test collection above.)

Relevance is clearly not black and white but has many shades of gray. This and the subjective nature of relevance calls for graded rather than binary relevance assessments (e.g., Sparck Jones & Rijsbergen, 1976). For example, INEX uses a four-point scale as defined by Kekäläinen and Järvelin (2002) (initially used by Sormunen, 2002):

- Highly relevant: The element discusses the topic of request exhaustively.
- Fairly relevant: The element discusses the topic of request, but not exhaustively.
- Marginally relevant: The element mentions the topic of request, but only in passing.
- Irrelevant: The element does not contain any information about the topic of request.

However, with the exception of some tracks, relevance assessments in TREC are binary (Ingwersen & Järvelin, 2005, p. 140).

Retrieval evaluation measures

A full discussion is found in Soergel & Golub (2015, under revision). The choice of measure for evaluating an automatic indexing tool depends on the purpose: computer-assisted indexing or fully automatic indexing. Apart from performance, aspects such as efficiency, utility, and cost have been used albeit more rarely.

In addition, failure analysis as well as analysis of success should be conducted, as much can be learned from it (Lancaster, 1968; Soergel, 1985, p. 390; Hersh et al. 1994; Saracevic 2008;). However, few studies include failure analysis because it is complex and requires much human effort.

Recommended approach for evaluating indexing quality through analyzing retrieval performance

The purpose of such evaluation is often to compare retrieval results based on indexing from different sources, for example, comparing available humanly assigned controlled terms, controlled or uncontrolled terms assigned by a computer program,

and free-text search. Retrieval results depend on the indexing and on the retrieval scoring method used; the two may interact. The test procedure described below is intended primarily for retrieval scenarios described in the introduction, but could be used with other retrieval approaches; the choice depends on the operational environment in which the indexing method(s) being evaluated (such as different automatic indexing programs) is/are to be used.

We suggest the following procedure

- Use a test collection of ~10,000 documents drawn from an operational collection with available controlled terms, covering several (three or more) subject areas. Index some or all of these documents with all of the indexing methods to be tested. 10,000 documents is a ball-park figure to make sure there are at least 20 relevant documents for three suitably chosen topics and enough irrelevant documents that are similar to relevant ones so that the ability of the retrieval system to discriminate can be tested.
- For each of the selected subject areas, choose a number of users; ideally, there would be equal numbers of end users, subject experts, and information professionals.
- The users conduct searches on several topics in their assigned subject area, some of the topics chosen by the user and some assigned:
    - one of the user topics should be an extensive search for an essay, thesis or similar, which would require finding an extensive list of documents; such a search would likely benefit from the index terms;
    - one of the user topics should be a minor, factual search for information; this may be less dependent on index terms;
    - the assigned topic should be selected to have at least 20 relevant documents in the test collection so that a reasonable number of documents is retrieved and available for analysis;
- Run the users' queries using all available evidence from all indexing sources to produce a "maximal" retrieved set and present results with query terms highlighted.
- Have users assess the relevance of each document found, ideally on a scale from 0 to 4, not relevant to highly relevant, and indicating the type of relevance (direct, indirect, context, comparison, or in more detail, see Huang & Soergel, 2013). In preparation, instruct the users/assessors in how to assess relevance in order to increase inter-rater consistency. Having different types of users assess relevance will cover several cognitively different representations (cf. the polyrepresentation concept, Ingwersen, 1996).
- Compute retrieval performance metrics for each individual indexing source (considering the documents found based on indexing from that source alone) and for selected combinations of indexing sources at different degrees of relevance.
- Perform log analysis, observe several people how they perform their tasks, and get feedback from the assessors through questionnaires and interviews in order to collect a wide variety of qualitative and quantitative data on the

contribution of assigned index terms and .of free-text terms to retrieval, considering also the effect of the user's query formulation

- Perform a detailed analysis of retrieval failures and retrieval successes, focusing on cases where indexing methods differ with respect to retrieving a relevant or irrelevant document.

Conclusion

Automatic subject indexing and classification can take over or assist with the creation of subject metadata, making it sustainable at large scale; enrich existing metadata; and establish more connections across resources.

While many state-of-the-art commercial software vendors and experimental researchers make claims about great success of automated subject assignment tools, hard evidence of their performance in operating information environments is scarce. Among major reasons for this is the fact that real-life studies which encompass complex aspects involved in evaluation of an operating system would be resource-heavy and time-consuming. Consequently, few such studies have been conducted and related research communities and publication channels accept laboratory studies to be as valid and reliable.

The paper discusses issues with laboratory-based information retrieval testing and proposes a framework for evaluation of indexing in general and automated subject assignment tools in particular as a guide to realistic evaluation. The paper addresses problems of aboutness and relevance assessments together with implications for the validity of the common gold standard approach. We argue that "gold standard" test collections should be thoroughly designed and built or adapted from an existing database such as a repository with careful planning and quality control and that they should be used judiciously with provisions for revision of relevance assessments in particular.

The proposed framework addresses different evaluation aspects, applying a triangulation of methods and exploring a variety of perspectives and contexts. Based on a thorough review of related research we present recommendations for

1 Evaluating indexing quality directly through assessment by an evaluator or by comparison with a gold standard.

2 Evaluating tools for computer-assisted indexing directly in the context of an indexing workflow.

3 Evaluating indexing quality indirectly through analyzing retrieval performance.

In order to make the framework more reliable, it should be further informed by empirical evidence from applying it, providing an opportunity to re-examine the proposed steps and their practical applicability and validity, as well as provide more insight into the degree to which such a complex approach would justify its cost. Likewise, it would be important to identify which use cases would benefit from this approach. When an evaluation is conducted to make important and expensive decisions, a thorough approach such as the one suggested in this paper is appropriate.

Finally, it is expected that much more research is required to develop appropriate experimental designs for such complex phenomena involving subject indexing and retrieval and information interaction in general. More serious scholarship needs to be devoted to evaluation in order to further our understanding of the value of automated subject assignment tools and to enable us to provide a fully informed input for their development and enhancement.

Acknowledgments

This work was in part funded by the JISC Information Environment Programme 2009-11 as the EASTER (Evaluating Automated Subject Tools for Enhancing Retrieval) project. We thank the anonymous reviewers whose comments helped significantly improve the paper.

References

20 Newsgroups DataSet. (1998). *The 4 universities data set*. Retrieved February 9, 2014, from http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html

Albright, J. B. (1979). *Some limits to subject retrieval from a large published index* (Doctoral dissertation). Urbana-Champaign, University of Illinois, Graduate School of Library Science.

Anderson, J. D., & Perez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval: Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing and Management 37*, 255-277.

Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., & Rogers, W. J. (2004). The NLM Indexing Initiative's Medical Text Indexer. In M. Fieschi et al. (Eds.), *Proceedings of the 11th World Congress On Medical Informatics* (pp. 268-272).

Bainbridge, D., & Witten, I. H. (2008). A Fedora librarian interface. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (pp. 407-416).

Beaulieu, M. (2003). Approaches to user-based studies in information seeking and retrieval: A Sheffield perspective. *Journal of Information Science 29(4)*, 239-248.

Belkin, N. et al. (2014). Task-based information retrieval. In Agosti, Maristella et al. (Eds.), *Evaluation Methodologies in Information Retrieval (Dagstuhl Seminar 13441)* (pp. 117-119).

Blandford, A., Keith, S., Connell, I., & Edwards, H. (2004). Analytical usability evaluation for digital libraries: A case study. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries* (pp. 27-36).

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology 54(10)*, 913-925.

Braschler, M., & Peters, C. (2004). Cross-language evaluation forum: Objectives, results, achievements. *Information Retrieval 7(1-2)*, 7-31.

Brenner, C. W., & Mooers, C. N. (1958). A case history of a Zatocoding information retrieval system. In R. S. Casey. J. W. Perry. M. M. Berry. & A. Kent (Eds.). *Punched cards* (pp. 340-356). New York: Reinhold.

Buckley, C., Dimmick, D., Soboroff, I., & Voorhees, E. (2007). Bias and the limits of pooling for large collections. *Information Retrieval 10(6),* 491-508.

Buckley, C., & Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *ACM SIGIR 2000 Proceedings* (pp. 33-40).

Chung, Y.-M., Pottenger, W. M., & Schatz, B. R. (1998). Automatic subject indexing using an associative neural network. In *Proceedings of the third ACM conference on Digital libraries* (pp. 59-68).

Cleverdon, C. W., Mills, J., & Keen, M. (1968). *Factors determining the performance of indexing systems*. Aslib Cranfield Research Project, Cranfield, England.

Colosimo, M.E., Morgan, A.A., Yeh, A.S., Colombe, J.B., & Hirschman L. (2005). Data preparation and interannotator agreement: BioCreAtIvE task 1B. *BMC Bioinformatics 6(Suppl 1)*.

Cooper, W. S. (1969). Is interindexer consistency a hobgoblin? *American Documentation*, 20, 268-278.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science 35*, 982-1003.

Fidel, R. (1994), User-centered indexing. *Journal of the American Society for Information Science*, 45: 572–576.

Golub, K. (2006a). Automated subject classification of textual web documents. *Journal of Documentation 62(3),* 350-371.

Golub, K. (2006b). Automated subject classification of textual web pages, based on a controlled vocabulary: challenges and recommendations. *New review of hypermedia and multimedia 12(1)*, 11-27.

Golub, K., & Lykke, M. (2009). Automated classification of web pages in hierarchical browsing. *Journal of Documentation 65(6),* 901–925.

Hagedorn, K. (2001). Extracting value from automated classification tools: The role of manual involvement and controlled vocabularies. Retrieved February 9, 2014, from http://argus-acia.com/white_papers/classification.html

Hersh, W. R., Hickam, D. H., Haynes, R. B., McKibbon, K. A. (1994). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of American Medical Information Association,* 51-60.

Hliaoutakis, A., Zervanou, K., & Petrakis, E. G. M. (2009). The AMTEx approach in the medical indexing and retrieval environment. *Data and Knowledge Engineering Journal 68(3),* 380-392.

Hosseini, M., Cox, I., Milic-Frayling, N., Kazai, G., Vinay, V., Baeza-Yates, R., & ... Silvestri, F. (2012). On aggregating labels from multiple crowd workers to infer relevance of documents. In *Proceedings of the 34th European Conference on IR Research (ECIR 2012)*.

Hripcsak, G., & Wilcox, A. (2002). Reference standards, judges, and comparison subjects: Roles for experts in evaluating system performance. *Journal of the American Medical Informatics Association 9(1)*, 1-15.

Huang, X., & Soergel, D. (2013). Functional relevance and inductive development of an e-retailing product information typology. *Information Research 18(2)*.

Iivonen, M. (1995). Consistency in the selection of search concepts and search terms. *Information Processing and Management 31(2),* 173-190.

Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, 52(1),* 3-50.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context.* Dordrecht: Springer.

International Organization for Standardization. (1985). *Documentation - methods for examining documents, determining their subjects, and selecting indexing terms : ISO 5963-1985*. Geneve: International Organization for Standardization.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, *16*(2), 138-178.

Kim, K. S. (2008). Effects of emotion control and task on web searching behavior. *Information Processing and Management 44(1)*, 373-385.

Kim, S., & Soergel, D. (2005). Selecting and measuring task characteristics as independent variables. *Proceedings ASIST* 68: 163–186.

Kekäläinen, J., & Järvelin, K. (2002). Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology 53(13),* 1120-1129.

Lalmas, M., & Tombros, A. (2007). Evaluating XML retrieval effectiveness at INEX. In *SIGIR Forum* (pp. 40-57).

Lancaster, F. W. (1968). *Evaluation of the MEDLARS Demand Search Service.* Bethesda: National Library of Medicine.

Lancaster, F. W. (2003). Indexing and abstracting in theory and practice. 3rd ed. Champaign: University of Illinois.

Lewis, D.D., Yang, Y., Rose, T., & Li, F. (2004). RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research 5,* 361-397.

Liu, J., & Belkin, N. J. (2015). Personalizing information retrieval for multi-session tasks: Examining the roles of task stage, task type, and topic knowledge on the interpretation of dwell time as an indicator of document usefulness. *Journal of the Association for Information Science and Technology, 66*, 58–81.

Lykke, M. & Eslau, A. G. (2010). Using thesauri in enterprise settings: Indexing or query expansion? In: Larsen, B. Schneider, J. W., Åström, F. & Schlemmer, B. (Eds), *The Janus Faced Scholar: a Festschrift in honour of Peter Ingwersen*. Copenhagen: Det Informationsvidenskabelige Akademi. 87-97.

Mai, J-E. (2001). Semiotics and Indexing: An Analysis of the Subject Indexing Process. *Journal of Documentation 57 (5)*, 591-622.

Markey, K. (1984). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library & Information Science Research, 6(2),* 155-177.

Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyphrase indexing. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (pp. 296-297).

Mladenic, D. (1998). Turning Yahoo into an automatic web-page classifier. In *Proceedings of the 13th European Conference on Artificial Intelligence* (pp. 473-474).

Moens, M.-F. (2000). *Automatic indexing and abstracting of document texts*. Boston: Kluwer.

National Library of Medicine. (2009). *Indexing Initiative*. Retrieved February 9, 2014, from http://ii.nlm.nih.gov/

National Library of Medicine. (2010). *Medical Text Indexer*. Retrieved February 9, 2014, from http://ii.nlm.nih.gov/mti.shtml

Oard, D. W., Soergel, D., Doermann, D., Huang, X., Murray, G. C., Wang, J., Ramabhadran, B., Franz, M., Gustman, S., Mayfield, J., Kharevych, L., & Strassel, S. (2004). Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (41-48).

Olson, H. A., & Boll, J. J. (2001). *Subject analysis in online catalogs*. 2nd ed. Englewood, CO: Libraries Unlimited.

Paynter, G. W. (2005). Developing practical automatic metadata assignment and evaluation tools for internet resources. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries* (pp. 7-11).

Plaunt, C., & Norgard, B. A. (1997). An association based method for automatic indexing with a controlled vocabulary. *Journal of the American Association for Information Science 49(10)*, 888-902.

Polfreman, M. & Grace, S. (2008). MetaTools stage 1 report: Test framework: A methodology for evaluating metadata generation tools. Retrieved February 9, 2014, from http://www.jisc.ac.uk/media/documents/programmes/reppres/metatoolsfinalreport.pdf

Purpura, S. & Hillard, D. (2006). Automated classification of congressional legislation. In *Proceedings of the 2006 international conference on digital government research* (pp. 219-225).

Ribeiro-Neto, B., Laender, A. H. F., & de Lima, L. R. S. (2001). An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Technology 52(5),* 391–401.

Roberts, D., & Souter, C. (2000). The automation of controlled vocabulary subject indexing of medical journal articles. *Aslib Proceedings 52(10)*, 384-401.

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management 17,* 69–76.

Roitblat, H. L., Kershaw, A., & Oot, P. (2010). Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology 61(1),* 70–80.

Rosenberg, V. (1971). Comparative evaluation of two indexing methods using judges. *Journal of the American Society for Information Science 22(4),* 251-259.

Ruiz, M. E., & Aronson, A. (2007). User-centered evaluation of the Medical Text Indexing (MTI) system. Retrieved February 9, 2014, from http://ii.nlm.nih.gov/resources/MTIEvaluation-Final.pdf

Ruiz, M. E., Aronson, A. R., & Hlava, M. (2008). Adoption and evaluation issues of automatic and computer aided indexing systems. In *Proceedings of the American Society for Information Science and Technology, 45(1)*, 1-4.

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science: Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology 58(13),* 1915-1933.

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science: Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology Volume 58(13)*, 2126-2144.

Saracevic, T. (2008). Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Library Trends 56(4)*, 763-783.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys 34(1)*, 1-47.

Silvester, J. P. (1997). Computer supported indexing: A history and evaluation of NASA's MAI system. In *Encylopedia of Library and Information Services 61(24),* 76-90.

Soergel, D. (1985). *Organizing information: Principles of data base and retrieval systems*. Orlando, FL: Academic Press.

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science 45(8),* 589-599.

Soergel, D., & Golub, K. (2015). Formal analysis of similarity measures with emphasis on consistency and performance measures in indexing and retrieval. *Under revision for Journal of the American Society for Information Science*.

Sormunen, E. (2002). Liberal relevance criteria of TREC – counting on negligible documents? In *Proceedings of the Twenty-Fifth Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 324-330).

Sparck Jones, K., & Rijsbergen, K. (1976). Information retrieval test collections. *Journal of Documentation 32,* 59-75.

Suomela, S., & Kekäläinen, J. (2006). User evaluation of ontology as query construction tool. *Information Retrieval 9(4),* 455-475.

Svarre, T.J., & Lykke, M. (2014). Experiences with automated categorization in e-government information retrieval. *Knowledge Organization*, 41(1). 76-84.

Svenonius, E. (2000). *The intellectual foundation of information organization*. Cambridge, Mass.: MIT Press.

Text REtrieval Conference: TREC. (2010). Retrieved February 9, 2014, from http://trec.nist.gov/

Tonkin, E., & Muller, H. (2008). Keyword and metadata extraction from pre-prints. In *Proceedings of the 12th International Conference on Electronic Publishing.* Pre-print retrieved February 9, 2014, from http://elpub.scix.net/data/works/att/030_elpub2008.content.pdf

Tsai, C.-F., McGarry, K., & Tait, J. (2006). Qualitative evaluation of automatic assignment of keywords to images. *Information Processing and Management 42*, 136–154.

Urbano, J., Marrero, M., & Martín, D. (2013). On the measurement of test collection reliability. In *Proceedings of ACM SIGIR* (pp. 393-402).

Venanzi, M., Guiver, J., Kazai, G., Kohli, P., & Shokouhi, M. (2014). Community-based bayesian aggregation models for crowdsourcing. Proceedings of the 23rd International Conference World Wide Web, 155.

Voorhees, E. M., & Harman, D. K. (2005). The Text REtrieval Conference. In Voorhees, E. M., & Harman, D. K. (Eds.), *TREC: Experiments and evaluation in information retrieval* (pp. 3-19).